Outlier identification for skewed and/or heavy-tailed unimodal multivariate distributions

Verardi, Vincenzo; Vermandele, Catherine

Published in:
Journal de la Société Française de Statistique

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Journal de la Société Française de Statistique
Vol. 157 No. 2 (2016)

Outlier identification for skewed and/or heavy-tailed unimodal multivariate distributions

Titre: Identification de valeurs extrêmes pour des distributions multivariées unimodales asymétriques et/ou à queues lourdes

Vincenzo Verardi[1] and Catherine Vermandele[2]

Abstract: In multivariate analysis, it is very difficult to identify outliers in case of skewed and/or heavy-tailed distributions. In this paper, we propose a very simple outlier identification tool that works with these types of distributions and that keeps the computational complexity low.

Résumé : L'identification de valeurs extrêmes s'avère particulièrement délicate en analyse multivariée lorsque la distribution sous-jacente est asymétrique et/ou à queues lourdes. Cet article présente une méthode d'identification extrêmement simple, bien adaptée à ce type de distribution et qui n'exige qu'une faible complexité calculatoire.

Keywords: outlier identification, skewed multivariate distribution, heavy-tailed multivariate distribution, Tukey *g*-and-*h* distribution
Mots-clés : identification de valeurs extrêmes, distribution multivariée asymétrique, distribution multivariée à queues lourdes, distribution de Tukey *g*-et-*h*
AMS 2000 subject classifications: 62E17, 62G07, 62G35, 62H10

1. Introduction

The detection of outliers in univariate and multivariate data is particularly tricky when the data are generated from skewed and/or heavy-tailed distributions. Indeed, most of the available outlier identification tools, such as the standard boxplot in the univariate case or, in the multivariate case, the robust Mahalanobis distances (based, for example, on the Minimum Covariance Determinant estimator of location and scatter), rely on the elliptical symmetry assumption.

In a recent paper, Hubert and Vandervieren (2008) propose a new outlier identification rule for skewed univariate data based on a so-called *adjusted boxplot*. Their idea is to modify the whiskers of the standard boxplot according to the degree of asymmetry in the data distribution, which can be robustly estimated by the medcouple. The expressions of the whiskers extremities of this adjusted boxplot were found from extensive simulations of a wide range of (moderately) skewed distributions and such that, in absence of contamination by outliers, approximately 0.7% of the observations lie outside the interval delimited by both whiskers (as it is the case for the standard boxplot and Gaussian data). By using this new univariate tool, Hubert and Van der

[1] University of Namur.
E-mail: vverardi@unamur.be
[2] Université libre de Bruxelles.
E-mail: vermande@ulb.ac.be

 ISSN: 2102-6238

Veeken (2008) propose a projection-based multivariate outlier detection method which does not rely on the elliptical symmetry assumption anymore. Their idea is based on the fact that, as stated previously by Stahel (1981) and Donoho (1982), a multivariate outlier is a univariate outlier in some direction of the space. As Stahel and Donoho, Hubert and Van der Veeken define the (global) *outlyingness* of a point $\mathbf{x}_i$ of the dataset $\mathscr{X}^{(n)} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ as the maximum of the outlyingness measures of this point along a large number of possible directions $\mathbf{a}$ of the space containing the data cloud. The outlyingness measure of $\mathbf{x}_i$ along a specific direction $\mathbf{a}$ is a robust measure of the distance of the projection of $\mathbf{x}_i$ on $\mathbf{a}$ with respect to the center of the projected data cloud. Hubert and Van der Veeken suggest to modify the measure originally proposed by Stahel and Donoho, in order to adjust for the skewness of the underlying distribution of the data; they propose a definition based on the bounds of the lower and upper whiskers of the adjusted boxplot of Hubert and Vandervieren (2008) associated to the projected cloud. A point $\mathbf{x}_i$ of the dataset $\mathscr{X}^{(n)}$ will finally be identified as an outlier if its global outlyingness is greater than the extremity of the upper whisker of the adjusted boxplot built from the global outlyingness measures of the n points of $\mathscr{X}^{(n)}$. One of the biggest drawbacks of this method is its computational complexity. With large datasets and high dimensions, numerous projections are needed and the computational time can become prohibitive. Additionally, as highlighted by Bruffaerts et al. (2014), the adjusted boxplot does not perform well in case of heavy-tailed and/or skewed distributions with bounded support. This pitfall naturally remains in the multivariate setup which limits substantially the class of distributions for which the method is attractive.

In this paper, we propose a new multivariate projection-based outlier identification tool that is more computationally efficient than the one proposed by Hubert and Van der Veeken (2008). Furthermore the proposed method is more general as in addition to skewness it also deals with heaviness of tails and bounded-support skewed-distributions.

Like Hubert and Van der Veeken (2008), we identify a point $\mathbf{x}_i$ of the dataset $\mathscr{X}^{(n)}$ as an outlier if its global outlyingness exceeds a certain cut-off value. But our method differs from the one of the previous authors in the way we define the outlyingness of a point $\mathbf{x}_i$ along a specific direction $\mathbf{a}$ of the space as well as in the manner of determining the bound to which we compare the global outlyingness measure of each point of $\mathscr{X}^{(n)}$. More precisely, for each considered direction of the space, we define an outlyingness measure along this direction that takes into account the skewness of the distribution of the distances between the projected points and their median but is much faster to compute than the adjusted outlyingness measure considered by Hubert and Van der Veeken (2008). Moreover, the bound allowing to identify outliers among the dataset $\mathscr{X}^{(n)}$ actually consists of an estimation of a specific upper quantile of the distribution of the global outlyingness measures of the observations $\mathbf{x}_i$. This distribution is unknown but, when it is unimodal, can be well adjusted — for the skewness as well as for the tails heaviness — by a specific distribution whose quantiles may be computed in an easy way. The underlying key idea of the procedure consists of the fact that a certain simple rank preserving monotonic transformation of the global outlyingness measures of the points $\mathbf{x}_i$ provides transformed measures whose distribution is very well approximated by a so-called *Tukey g-and-h distribution*.

The structure of the paper is as follows: In Section 2, we present the *adjusted* outlier identification method of Hubert and Van der Veeken (2008) and, in Section 3, we introduce the Tukey g-and-h distribution. In Section 4, we detail the proposed outlier identification procedure.

Section 5 is devoted to empirical results. Firstly, we present a generated example that illustrates

 ISSN: 2102-6238

the methodology in absence as well as in presence of contamination of the data by outliers. Secondly, we consider bivariate samples generated from six unimodal distributions that strongly differ in terms of skewness and tails heaviness, and we study the quality of the adjustment of the (transformed) global outlyingness measures distribution from a Tukey g-and-h distribution.

In Section 6, we give some simulations results about the general performances of the new outlier identification method. Three different sets of simulations are presented: The first two sets allow us to study the sensitivity and specificity of the method for various sample sizes and data dimensions, and for a wide variety of distributions of the data. With the third set of simulations, we compare the new outlier identification method with its closest competitor — the method of Hubert and Van der Veeken (2008) — in terms of sensitivity as well as in terms of computational complexity and time.

We present an empirical application in Section 7 and conclude in Section 8.

2. Projection-based multivariate outlier detection methods

2.1. The Stahel-Donoho outlyingness measure

Consider a p-dimensional sample $\mathscr{X}^{(n)} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ with $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^t$. In this multivariate setup, an outlier is a point $\mathbf{x}_i$ that lies far away from the bulk of the data in any direction. It can therefore be seen as an outlier among the projections of the data points on a particular direction of the space $\mathbb{R}^p$.

Following this idea, Stahel (1981) and Donoho (1982) proposed to measure the outlyingness of an observation by considering the projection of this observation on the direction of the space along which the observation is most outlying. More precisely, given a direction $\mathbf{a} \in \mathbb{R}^p$ with $\|\mathbf{a}\| = 1$, denote by $\mathscr{X}_{\mathbf{a}}^{(n)} = \{\mathbf{x}_1^t\mathbf{a}, \ldots, \mathbf{x}_n^t\mathbf{a}\}$ the projection of the dataset $\mathscr{X}^{(n)}$ upon $\mathbf{a}$. Let $\widehat{\mu}$ and $\widehat{\sigma}$ be robust univariate location and dispersion statistics, e.g., the median and MAD, respectively. The outlyingness with respect to $\mathscr{X}^{(n)}$ of a point $\mathbf{x} \in \mathbb{R}^p$ along $\mathbf{a}$ is defined as

$$\mathrm{SDO}_{\mathbf{a}}\left(\mathbf{x}; \mathscr{X}^{(n)}\right) = \frac{\left|\mathbf{x}^t\mathbf{a} - \widehat{\mu}\left(\mathscr{X}_{\mathbf{a}}^{(n)}\right)\right|}{\widehat{\sigma}\left(\mathscr{X}_{\mathbf{a}}^{(n)}\right)}. \tag{1}$$

The (global) *Stahel-Donoho outlyingness* of $\mathbf{x}$ with respect to $\mathscr{X}^{(n)}$ is then given by

$$\mathrm{SDO}\left(\mathbf{x}; \mathscr{X}^{(n)}\right) = \sup_{\mathbf{a} \in \mathscr{S}_p} \mathrm{SDO}_{\mathbf{a}}\left(\mathbf{x}; \mathscr{X}^{(n)}\right), \tag{2}$$

with $\mathscr{S}_p = \{\mathbf{a} \in \mathbb{R}^p : \|\mathbf{a}\| = 1\}$ [1]. From now on, we will denote $\mathrm{SDO}\left(\mathbf{x}_i; \mathscr{X}^{(n)}\right) = \mathrm{SDO}_i$ ($i = 1, \ldots, n$) for $\mathbf{x}_i \in \mathscr{X}^{(n)}$.

Note that, in practice, the global outlyingness measure SDO_i ($i = 1, \ldots, n$) cannot be computed exactly since it is impossible to project the observation $\mathbf{x}_i$ on *all* vectors $\mathbf{a} \in \mathscr{S}_p$. Hence, we must settle for an approximate value of SDO_i by restricting ourselves to a finite set $\widehat{\mathscr{S}_p}$ of randomly

[1] Note that, if $\widehat{\mu}$ and $\widehat{\sigma}$ are the mean and the standard deviation, then $\mathrm{SDO}\left(\mathbf{x}; \mathscr{X}^{(n)}\right) = d(\mathbf{x}, \overline{\mathbf{x}}; \mathbf{S}) = \sqrt{(\mathbf{x} - \overline{\mathbf{x}})^t \mathbf{S}^{-1} (\mathbf{x} - \overline{\mathbf{x}})}$, where $\overline{\mathbf{x}}$ and $\mathbf{S}$ are the mean and covariance matrix of the sample $\mathscr{X}^{(n)}$.

 ISSN: 2102-6238

selected directions. Many simulations have shown that considering $m = 250p$ directions yields a good balance between the computational feasibility and the quality of the obtained approximate solution (see Maronna and Yohai, 1995 for an extensive discussion).

If the data $\mathbf{x}_i$ $(i = 1, \ldots, n)$ are normally distributed, the global outlyingness measures SDO_i $(i = 1, \ldots, n)$ are asymptotically χ^2_p distributed (Maronna and Yohai, 1995). Hence, it is usual to identify a multivariate observation $\mathbf{x}_i$ as an outlier if its outlyingness measure SDO_i exceeds the $(1-\alpha)$-quantile of the χ^2_p distribution, for an arbitrary, small probability level α.

This outlier identification procedure based on the Stahel-Donoho outlyingness measures suffers from two major problems. Firstly, when the data $\mathbf{x}_i$ $(i = 1, \ldots, n)$ are not Gaussian, the distribution of the outlyingness measures SDO_i is in general unknown (but typically right-skewed as they are bounded by zero). In that case, the outlier detection rule based on the $(1-\alpha)$-quantile of the χ^2_p distribution risks to be unvalid.

Secondly, as stated by Hubert and Van der Veeken, the Stahel-Donoho outlyingness (1) does not account for any possible skewness of the distribution of the projected dataset $\mathscr{X}_{\mathbf{a}}^{(n)}$, since it assumes that the scale on the lower side of the median $\widehat{\mu}(\mathscr{X}_{\mathbf{a}}^{(n)})$ is the same as the scale on the upper side. Hence, the outlyingness measure of Stahel and Donoho is only suited for elliptical symmetric data.

In order to deal with these two problems, Hubert and Van der Veeken (2008) have proposed the *adjusted* outlier detection method described hereafter.

2.2. *The adjusted outlier detection method of Hubert and Van der Veeken (2008)*

First of all, Hubert and Van der Veeken (2008) propose to modify the Stahel-Donoho outlyingness measures $\mathrm{SDO}_{\mathbf{a}}\left(\mathbf{x}; \mathscr{X}^{(n)}\right)$ to make them suitable for skewed data. Their definition involves the extremities of the whiskers of the adjusted boxplot introduced by Hubert and Vandervieren (2008). The latter authors suggest to modify the whiskers of the classical boxplot according to the degree of asymmetry of the data. More precisely, they propose to define, for the univariate statistical series $\{y_1, \ldots, y_n\}$, an *adjusted* boxplot for which the extremities of the whiskers correspond to the bounds of the following interval:

$$\begin{cases} \left[Q_{0.25} - 1.5e^{-4\,\mathrm{MC}}\mathrm{IQR}\,; Q_{0.75} + 1.5e^{3\,\mathrm{MC}}\mathrm{IQR}\right] & \text{if } \mathrm{MC} \geq 0 \\ \left[Q_{0.25} - 1.5e^{-3\,\mathrm{MC}}\mathrm{IQR}\,; Q_{0.75} + 1.5e^{4\,\mathrm{MC}}\mathrm{IQR}\right] & \text{if } \mathrm{MC} < 0, \end{cases} \tag{3}$$

where $Q_{0.25}$ and $Q_{0.75}$ are the first and third quartiles of the series $\{y_1, \ldots, y_n\}$, $\mathrm{IQR} = Q_{0.75} - Q_{0.25}$ is its interquartile range, and MC stands for the medcouple [2] of the y_i's, which is a robust measure of skewness. This measure is bounded between -1 and 1; the medcouple is equal to zero when the observed distribution of the y_i's is symmetric, whereas a positive (*resp.* negative) value of MC corresponds to a right (*resp.* left) tailed distribution. These expressions for the extremities of the whiskers of the adjusted boxplot have been found by simulating a wide range of skewed distributions and looking for the interval which leaves 0.7% [3] of the observations outside its bounds when no outlier contamination is present.

[2] See Brys et al. (2004).

[3] This percentage corresponds to the theoretical proportion of observations that, in a Gaussian dataset, will lie outside the interval $[Q_{0.25} - 1.5\ \mathrm{IQR}\,; Q_{0.75} + 1.5\ \mathrm{IQR}]$ whose bounds define the extremities of the lower and upper whiskers of the classical boxplot.

 ISSN: 2102-6238

Hubert and Van der Veeken (2008) define the *adjusted* outlyingness of a point $\mathbf{x} \in \mathbb{R}^p$ with respect to $\mathscr{X}^{(n)}$ along $\mathbf{a}$ as

$$\mathrm{AO}_{\mathbf{a}}\left(\mathbf{x}; \mathscr{X}^{(n)}\right) = \begin{cases} \frac{\mathbf{x}^t\mathbf{a} - Q_{0.5}(\mathscr{X}_{\mathbf{a}}^{(n)})}{u_2(\mathscr{X}_{\mathbf{a}}^{(n)}) - Q_{0.5}(\mathscr{X}_{\mathbf{a}}^{(n)})} & \text{if } \mathbf{x}^t\mathbf{a} \geq Q_{0.5}(\mathscr{X}_{\mathbf{a}}^{(n)}) \\ \frac{Q_{0.5}(\mathscr{X}_{\mathbf{a}}^{(n)}) - \mathbf{x}^t\mathbf{a}}{Q_{0.5}(\mathscr{X}_{\mathbf{a}}^{(n)}) - u_1(\mathscr{X}_{\mathbf{a}}^{(n)})} & \text{if } \mathbf{x}^t\mathbf{a} < Q_{0.5}(\mathscr{X}_{\mathbf{a}}^{(n)}) \end{cases},$$

where $Q_{0.5}(\mathscr{X}_{\mathbf{a}}^{(n)})$ is the median of the projected dataset $\mathscr{X}_{\mathbf{a}}^{(n)}$, and $u_1(\mathscr{X}_{\mathbf{a}}^{(n)})$ and $u_2(\mathscr{X}_{\mathbf{a}}^{(n)})$ are respectively the lower and upper whiskers extremities of the *adjusted* boxplot associated with $\mathscr{X}_{\mathbf{a}}^{(n)}$, that is, the lower and upper bounds of the interval (3) computed for $\mathscr{X}_{\mathbf{a}}^{(n)}$. This definition takes into account the fact that the distribution of the projected points $\mathbf{x}_i^t\mathbf{a}$ may be skewed and that this skewness will induce a difference between the scale of the part of the distribution on the left of the median and the scale of the part of the distribution on the right of the median.

The *adjusted* (global) *outlyingness of* $\mathbf{x}$ *with respect to* $\mathscr{X}^{(n)}$ is given by

$$\mathrm{AO}\left(\mathbf{x}; \mathscr{X}^{(n)}\right) = \sup_{\mathbf{a} \in \widehat{\mathscr{S}_p}} \mathrm{AO}_{\mathbf{a}}\left(\mathbf{x}; \mathscr{X}^{(n)}\right). \tag{4}$$

Hubert and Van der Veeken (2008) suggest then to identify a multivariate observation $\mathbf{x}_i$ as an outlier if its adjusted outlyingness $\mathrm{AO}_i = \mathrm{AO}\left(\mathbf{x}_i; \mathscr{X}^{(n)}\right)$ exceeds the upper whisker of the adjusted boxplot associated with $\{\mathrm{AO}_1, \ldots, \mathrm{AO}_n\}$.

This approach has the advantage to be distribution-free since it does not assume any particular underlying skewed distribution of the data (only unimodality).

However, the computational complexity of this procedure is substantial since for each of the $m = 250p$ considered directions $\mathbf{a} \in \mathbb{R}^p$, a medcouple (that has a computational complexity of $O(n \log n)$) has to be estimated. The computational complexity of the method is therefore $O(np \log n)$. With large datasets and high dimensions, computing time can become prohibitive [4].

Moreover, the outlier identification method of Hubert and Van der Veeken (2008) presents some limitations related to the use of the adjusted boxplot of Hubert and Vandervieren (2008). As already mentioned, the expressions of the adjusted whiskers have been determined by simulating a wide range of skewed distributions and looking for the interval which leaves 0.7% of the observations outside its bounds when no outlier contamination is present. More precisely, in their simulations, Hubert and Vandervieren only considered distributions with a medcouple smaller than 0.6 and, consequently, the theoretical detection rate of atypical observations associated with the adjusted boxplot risks to be quite different from 0.7% when the underlying distribution is severely skewed. Moreover, if one is interested in fixing a theoretical detection rate of atypical observations different than the standard rate of 0.7%, it is necessary to re-run all the extensive simulations in order to find the tuning factors pre-multiplying IQR in the expressions (3) of the whiskers extremities. Finally, the adjusted boxplot does not take into account the tail heaviness of the distribution of the y_i's, and, in particular, does not provide an adequate detection rate of atypical observations if the distribution of the AO_i's is heavy-tailed (see Bruffaerts et al., 2014).

To overcome these problems, we propose:

[4] In Section 6.3, we present a comparison, in terms of computational time, between the adjusted outlier detection method of Hubert and Van der Veeken and the new method proposed in this paper.

 ISSN: 2102-6238

1. a modified definition of the outlyingness of a point $\mathbf{x} \in \mathbb{R}^p$ with respect to $\mathscr{X}^{(n)}$ along a direction $\mathbf{a}$ that takes into account the skewness of the distribution of the projected points $\mathbf{x}_i^t\mathbf{a}$ but does not require to determine the medcouple of this distribution and is very fast to compute;
2. a new outlier identification rule based on the global outlyingness measures of the observations $\mathbf{x}_i$ $(i = 1,\ldots,n)$ that (i) has a low computational cost, (ii) allows the user to choose the detection rate of atypical observations theoretically reached in absence of contamination of the sample $\mathscr{X}^{(n)}$ by outliers, and (iii) respects this theoretical detection rate whatever the skewness level and the right tail heaviness of the distribution of the global outlyingness measures may be. The key idea consists of applying a very simple rank preserving transformation on the global outlyingness measures of the $\mathbf{x}_i$'s and of adjusting the distribution of these transformed measures by a so-called Tukey g-and-h distribution.

3. The Tukey g-and-h distribution

In the late 70's, Tukey (1977) introduced a new family of distributions, called Tukey g-and-h distributions, based on elementary transformations of the standard normal.

For g and $h \in \mathbb{R}$, consider the one-to-one monotone function $\tau_{g,h}(\cdot)$ defined on $\mathbb{R}$ as follows: For $g \neq 0$,

$$\tau_{g,h}(z) = \frac{1}{g}\left[\exp(gz) - 1\right]\exp\left(hz^2/2\right)$$

and, for $g = 0$,

$$\tau_{0,h}(z) = \lim_{g\to 0}\tau_{g,h}(z) = z\exp\left(hz^2/2\right).$$

Let Z be a random variable with standard normal distribution $N(0,1)$. Then, for $A \in \mathbb{R}$ and $B \in \mathbb{R}_0^+$, the random variable Y defined through the transformation

$$Y = A + B\tau_{g,h}(Z)$$

is said to have a Tukey g-and-h distribution with location parameter A and scale parameter B:

$$Y \sim T_{g,h}(A,B).$$

The parameter g controls the direction and the degree of skewness [5], while h controls the tail thickness (or elongation) of the distribution (see Hoaglin et al., 1985). The family of $T_{g,h}(A,B)$ distributions is very flexible and approximates well many commonly used distributions (Martinez and Iglewicz, 1984; MacGillivray, 1992; Jiménez and Arunachalam, 2011).

Different procedures for the estimation of the parameters of the $T_{g,h}(A,B)$ distribution have been proposed in the literature (see Hoaglin et al., 1985; Jiménez and Arunachalam, 2011; Mahbubul et al., 2008; Xu et al., 2014; Xu and Genton, 2015). Relying on Hoaglin et al. (1985) and Jiménez and Arunachalam (2011), we propose here to use the simplified and robust estimators defined hereafter (the justification of the use of these estimators is detailed in Appendix 1).

[5] $g = 0$ corresponds to a symmetric distribution; $g > 0$ yields a right-skewed distribution while $g < 0$ gives a left-skewed distribution.

 ISSN: 2102-6238

Let $Z \sim N(0,1)$, $Y = A + B\tau_{g,h}(Z) \sim T_{g,h}(A,B)$ and $\mathscr{Y}^{(n)} = \{y_1, \ldots, y_n\}$ be a series of n independent realizations of Y. For $\nu \in (0,1)$, let us denote by z_ν, y_ν and $Q_\nu\left(\mathscr{Y}^{(n)}\right)$ the quantile of order ν of the $N(0,1)$ distribution, of the $T_{g,h}(A,B)$ distribution and of the series $\mathscr{Y}^{(n)}$, respectively. Then:

(i) The location parameter A is simply estimated by the empirical median $Q_{0.5}\left(\mathscr{Y}^{(n)}\right)$ of the data.

(ii) A natural estimate of the parameter g is given by

$$\widehat{g}_\nu = \frac{1}{z_\nu} \ln\left(\frac{\mathrm{UHS}_\nu(\mathscr{Y}^{(n)})}{\mathrm{LHS}_\nu(\mathscr{Y}^{(n)})}\right)$$

for any fixed order $\nu \in (0.5,1)$, where $\mathrm{UHS}_\nu(\mathscr{Y}^{(n)})$ and $\mathrm{LHS}_\nu(\mathscr{Y}^{(n)})$ are the ν-th upper and lower half spread of the series $\mathscr{Y}^{(n)}$:

$$\begin{aligned}\mathrm{UHS}_\nu(\mathscr{Y}^{(n)}) &= Q_\nu(\mathscr{Y}^{(n)}) - Q_{0.5}(\mathscr{Y}^{(n)}),\\ \mathrm{LHS}_\nu(\mathscr{Y}^{(n)}) &= Q_{0.5}(\mathscr{Y}^{(n)}) - Q_{1-\nu}(\mathscr{Y}^{(n)}).\end{aligned}$$

We propose to choose $\nu = 0.9$, so that the estimator $\widehat{g}_\nu$ of g has a breakdown point [6] of 10%.

(iii) Let us consider the empirical interquartile range $\mathrm{IQR}(\mathscr{Y}^{(n)})$, together with the skewness measure $\mathrm{SK}(\mathscr{Y}^{(n)})$ and the kurtosis (elongation) measure $\mathrm{T}(\mathscr{Y}^{(n)})$ of $\mathscr{Y}^{(n)}$ defined as follows:

$$\begin{aligned}\mathrm{IQR}(\mathscr{Y}^{(n)}) &= Q_{0.75}(\mathscr{Y}^{(n)}) - Q_{0.25}(\mathscr{Y}^{(n)}),\\ \mathrm{SK}(\mathscr{Y}^{(n)}) &= \frac{Q_{0.9}(\mathscr{Y}^{(n)}) + Q_{0.1}(\mathscr{Y}^{(n)}) - 2Q_{0.5}(\mathscr{Y}^{(n)})}{Q_{0.9}(\mathscr{Y}^{(n)}) - Q_{0.1}(\mathscr{Y}^{(n)})},\\ \mathrm{T}(\mathscr{Y}^{(n)}) &= \frac{Q_{0.9}(\mathscr{Y}^{(n)}) - Q_{0.1}(\mathscr{Y}^{(n)})}{Q_{0.75}(\mathscr{Y}^{(n)}) - Q_{0.25}(\mathscr{Y}^{(n)})}.\end{aligned}$$

Let us also define the function

$$\varphi(s,t) = 0.6817766 + 0.0534282\, s + 0.1794771\, t - 0.0059595\, t^2$$

for $s,t \in \mathbb{R}$. Then, as explained in Appendix 1, the scale parameter B can be estimated by

$$\widehat{B} = \frac{0.7413\ \mathrm{IQR}(\mathscr{Y}^{(n)})}{\varphi(\mathrm{SK}(\mathscr{Y}^{(n)}), \mathrm{T}(\mathscr{Y}^{(n)}))}\ .$$

Since $\widehat{B}$ is defined on the basis of the empirical quantiles of order 0.10, 0.25, 0.5, 0.75 and 0.90, it has a breakdown point of 10%.

[6] Intuitively, the breakdown point of an estimator is the maximal proportion of outlying observations (e.g. arbitrarily large observations) the estimator can handle before breaking down (e.g. giving an arbitrary value). The higher the breakdown point of an estimator, the more robust it is.

 ISSN: 2102-6238

(iv) Finally, a natural estimate of h is, for any fixed $\nu \in (0.5, 1)$,

$$\widehat{h}_\nu = \frac{2}{z_\nu^2} \ln\left(-\widehat{g}_\nu \frac{Q_\nu(\mathscr{Y}^{*(n)}) Q_{1-\nu}(\mathscr{Y}^{*(n)})}{Q_\nu(\mathscr{Y}^{*(n)}) + Q_{1-\nu}(\mathscr{Y}^{*(n)})} \right)$$

where $\mathscr{Y}^{*(n)} = \{y_1^*, \ldots, y_n^*\}$ with

$$y_i^* = \frac{y_i - \widehat{A}}{\widehat{B}}, \qquad i = 1, \ldots, n.$$

Once again, we take $\nu = 0.9$ in order to ensure a breakdown point of 10% for the estimator of h.

4. The proposed outlier detection method

4.1. *The asymmetrical outlyingness measures*

As Hubert and Van der Veeken (2008), we propose to modify the Stahel-Donoho outlyingness measure $\mathrm{SDO}_{\mathbf{a}}\left(\mathbf{x}; \mathscr{X}^{(n)}\right)$ to take into account the asymmetry of the distribution of the projected points $\mathbf{x}_i^t\mathbf{a}$. We define the *asymmetrical* outlyingness with respect to $\mathscr{X}^{(n)}$ of a point $\mathbf{x} \in \mathbb{R}^p$ along $\mathbf{a}$ as follows:

$$\mathrm{ASO}_{\mathbf{a}}\left(\mathbf{x}; \mathscr{X}^{(n)}\right) = \begin{cases} \frac{\mathbf{x}^t\mathbf{a} - Q_{0.5}(\mathscr{X}_{\mathbf{a}}^{(n)})}{2c\left[Q_{0.75}(\mathscr{X}_{\mathbf{a}}^{(n)}) - Q_{0.5}(\mathscr{X}_{\mathbf{a}}^{(n)})\right]} & \text{if } \mathbf{x}^t\mathbf{a} \geq Q_{0.5}(\mathscr{X}_{\mathbf{a}}^{(n)}) \\ \frac{Q_{0.5}(\mathscr{X}_{\mathbf{a}}^{(n)}) - \mathbf{x}^t\mathbf{a}}{2c\left[Q_{0.5}(\mathscr{X}_{\mathbf{a}}^{(n)}) - Q_{0.25}(\mathscr{X}_{\mathbf{a}}^{(n)})\right]} & \text{if } \mathbf{x}^t\mathbf{a} < Q_{0.5}(\mathscr{X}_{\mathbf{a}}^{(n)}) \end{cases}, \tag{5}$$

where $Q_{0.5}(\mathscr{X}_{\mathbf{a}}^{(n)})$, $Q_{0.25}(\mathscr{X}_{\mathbf{a}}^{(n)})$ and $Q_{0.75}(\mathscr{X}_{\mathbf{a}}^{(n)})$ are respectively the median, the first quartile and the third quartile of the projected dataset $\mathscr{X}_{\mathbf{a}}^{(n)}$, and $c = 1/(z_{0.75} - z_{0.25}) = 1/1.34898 = 0.7413$ is a constant factor ensuring, in the Gaussian case, the consistency of the modified scale estimator c IQR for the scale parameter σ (the standard deviation). The *asymmetrical* (global) *outlyingness of* $\mathbf{x}$ *with respect to* $\mathscr{X}^{(n)}$ is then given by

$$\mathrm{ASO}\left(\mathbf{x}; \mathscr{X}^{(n)}\right) = \sup_{\mathbf{a} \in \widehat{\mathscr{S}_p}} \mathrm{ASO}_{\mathbf{a}}\left(\mathbf{x}; \mathscr{X}^{(n)}\right). \tag{6}$$

In the standardization of $\mathrm{ASO}_{\mathbf{a}}\left(\mathbf{x}; \mathscr{X}^{(n)}\right)$, we consider scale measures of the distribution of $\mathscr{X}_{\mathbf{a}}^{(n)}$ relying on the length of the right or left part of the interquartile interval [7]. These scale measures both have a breakdown point of 25%. Note that a more robust alternative to define the asymmetrical outlyingness measures would be to standardize the deviation between $\mathbf{x}^t\mathbf{a}$ and $Q_{0.5}(\mathscr{X}_{\mathbf{a}}^{(n)})$ by considering, for instance, the Q_n coefficient of dispersion of Rousseeuw and Croux (1993). This coefficient does not assume the symmetry of the distribution of $\mathscr{X}_{\mathbf{a}}^{(n)}$ and has a breakdown point of 50%. However, since the next steps of the outlier detection procedure will not guarantee such a high breakdown point for the complete methodology, we prefer to use the half interquartile range that is very fast to compute.

[7] Note that $Q_{0.75}(\mathscr{X}_{\mathbf{a}}^{(n)}) - Q_{0.5}(\mathscr{X}_{\mathbf{a}}^{(n)}) = \mathrm{UHS}_{0.75}(\mathscr{X}_{\mathbf{a}}^{(n)})$ and $Q_{0.5}(\mathscr{X}_{\mathbf{a}}^{(n)}) - Q_{0.25}(\mathscr{X}_{\mathbf{a}}^{(n)}) = \mathrm{LHS}_{0.75}(\mathscr{X}_{\mathbf{a}}^{(n)})$.

 ISSN: 2102-6238

4.2. *The outlier identification rule*

Let us now consider the asymmetrical outlyingness measures $\text{ASO}_i = \text{ASO}\left(\mathbf{x}_i; \mathscr{X}^{(n)}\right)$ for $i = 1,\ldots,n$. The guidelines of the new outlier identification rule are the following:

1. Standardize these outlyingness measures in order to obtain new values belonging to the open interval $(0,1)$: for $i = 1,\ldots,n$, compute

$$\widetilde{\text{ASO}}_i = \frac{\text{ASO}_i}{\min_{1\leq j\leq n}(\text{ASO}_j) + \max_{1\leq j\leq n}(\text{ASO}_j)}.$$

2. Consider the inverse normal (also called probit) transformation: for $i = 1,\ldots,n$,

$$w_i = \Phi^{-1}\left(\widetilde{\text{ASO}}_i\right)$$

   where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal. Note that this is a monotonic transformation which preserves the ranks.

3. Adjust the distribution of the values w_i $(i = 1,\ldots,n)$ by the Tukey $T_{\widehat{g},\widehat{h}}(\widehat{A},\widehat{B})$ distribution, where $\widehat{A}$, $\widehat{B}$, $\widehat{g}$ and $\widehat{h}$ are the estimates of the location, scale, skewness and tails heaviness parameters defined in Section 3, with the series $\mathscr{Y}^{(n)}$ corresponding here to the series $\mathscr{W}^{(n)} = \{w_1,\ldots,w_n\}$.

4. Determine the quantile $\xi_{1-\alpha}$ of order $1-\alpha$ $(\alpha \in (0,0.5))$ of the $T_{\widehat{g},\widehat{h}}(\widehat{A},\widehat{B})$ distribution specified in the previous step, where α corresponds to the desired detection rate of atypical values in absence of contamination by outliers:

$$\xi_{1-\alpha} = \widehat{A} + \widehat{B}\tau_{\widehat{g},\widehat{h}}(z_{1-\alpha}).$$

   Let $\mathscr{I} = \{i = 1,\ldots,n | w_i > \xi_{1-\alpha}\}$ be the set of indices of the values w_i that are detected as atypically large in the set $\mathscr{W}^{(n)} = \{w_i,\ldots,w_n\}$; then the outlyingness measures ASO_i with $i \in \mathscr{I}$ are considered as atypical measures [8] among $\text{ASO}_1,\ldots,\text{ASO}_n$, and the observations $\mathbf{x}_i$ with $i \in \mathscr{I}$ are identified as outliers in the initial dataset $\mathscr{X}^{(n)}$.

Two remarks may complete the description of the outlier identification procedure:

(i) By inverting the transformations described in the first two steps, it is possible to come up with a *cut-off point* which is explicitly associated with the asymmetrical outlyingness measures ASO_i $(i = 1,\ldots,n)$. More precisely, the detection bound $\xi_{1-\alpha}$ computed in Step 4 leads to the following detection bound $B_+(\alpha)$ for the original $\text{ASO}_1,\ldots,\text{ASO}_n$:

$$B_+(\alpha) = \Phi(\xi_{1-\alpha})\left[\min_{1\leq j\leq n}(\text{ASO}_j) + \max_{1\leq j\leq n}(\text{ASO}_j)\right]. \tag{7}$$

(ii) Using the same logic as in the previous remark, we can easily obtain an estimation of the entire density function of the original asymmetrical outlyingness measures ASO_i $(i = 1,\ldots,n)$. In practice, we may proceed as follows. For $k = 1,\ldots,n$, we determine the

[8] Only observations $\mathbf{x}_i$ providing large outlyingness measures ASO_i, and hence large values w_i, are candidates to be outliers. It is then sufficient to consider the upper quantile $\xi_{1-\alpha}$ to define the outlier identification rule.

 ISSN: 2102-6238

quantiles ξ_{ν_k} of orders $\nu_k = k/(n+1)$ of the $T_{\widehat{g},\widehat{h}}(\widehat{A},\widehat{B})$ distribution specified in Step 3. We then apply the inverse transformation (7) in which we replace $\xi_{1-\alpha}$ by the quantiles ξ_{ν_k} in order to get the quantiles d_{ν_k} of orders ν_k on the scale of the asymmetrical outlyingness measures ASO_i. The kernel estimated density of the quantiles d_{ν_k} provides an estimation of the density of the ASO_i. We call it the *Tukey-based* density estimation.

The rationale for the transformations applied on the outlyingness measures in the first two steps of the procedure is the following. In Step 1, we bound these measures between 0 and 1: Our approach is comparable to the logic of dividing the ranks of some observations by the minimal rank (equal to 1) plus the maximal rank (equal to n) when we want to work with rank-based scores. The main difference is that our transformed asymmetrical outlyingness measures $\widetilde{\mathrm{ASO}}_i$ $(i = 1,\ldots,n)$ are in general not uniformly distributed on the interval $(0,1)$. Indeed, our transformation actually preserves both the skewness and the tails heaviness of the distribution of the initial ASO_i's. In Step 2, we consider the inverse normal transformation. If the $\widetilde{\mathrm{ASO}}_i$'s would be uniformly distributed on $(0,1)$, the w_i's would be normallly distributed. Since the $\widetilde{\mathrm{ASO}}_i$'s are not uniformly distributed on $(0,1)$, the distribution of the w_i's is not the standard normal distribution but a transformation of the normal distribution which allows for skewness and tails heaviness. Since the Tukey g-and-h distribution also appears as a transformation of the normal distribution allowing for a very large flexibility in skewness and tails heaviness, it appears as an appropriate distribution to adjust properly the observed distribution of the w_i's, whatever is the underlying unimodal distribution of the original data $\mathbf{x}_i$ $(i = 1,\ldots,n)$.

Note that, for a large variety of (unimodal and smooth) distributions of the data $\mathbf{x}_i$, the resulting distribution of the asymmetrical outlyingness measures ASO_i should be directly — without the preliminary transformations of Step 1 and Step 2 — adequately adjusted by a Tukey g-and-h distribution. But if the distribution of the ASO_i's is not smooth everywhere — if, for instance, the distribution of the ASO_i's looks like a triangular distribution — the quality of its adjustment by a Tukey distribution risks to be poor. In such a case, to transform the ASO_i's as indicated in Step 1 and Step 2 has the real advantage to provide transformed outlyingness measures w_i that have a smoother distribution, more properly adjusted by a Tukey g-and-h distribution.

Of course, it would have been possible to consider other distributions than the Tukey g-and-h distributions to adjust the distribution of the (transformed) asymmetrical outlyingness measures: we could use, fo example, the quite popular SAS-normal distributions of Jones and Pewsey (2009) or, more generally, any other transformation of the normal distribution that has been proposed in the literature (see Ley, 2015). Our choice of the $T_{g,h}(A,B)$ distribution is motivated by: (i) the great flexibility of this type of distribution, in terms of skewness and tails weight; (ii) the fact that we can estimate its parameters in a simple and robust way. Recall here that the estimation procedure we propose relies exclusively on percentiles 10, 25, 50, 75 and 90, which means that the breakdown point of the estimators of A, B, g and h is equal to 10%. In other terms, the distribution of the w_i's can be properly adjusted even in presence of (at most) 10% of outliers among the transformed outlyingness measures. This property of robustness is, of course, crucial to ensure the validity of our outlier detection procedure.

Let us finally note that the method proposed here works well in high dimensions as well. It is even possible to use it when $p > n$ if the type of projection used is some projection pursuit algorithm. We however did not consider high dimensions here.

 ISSN: 2102-6238

5. Empirical results

5.1. *Illustrative example*

To illustrate the methodology, we generate 1000 bivariate observations from two independent chi-squared distributions with ten degrees of freedom (χ^2_{10}). We first generate a clean sample and then contaminate it by replacing 5% of the observations with the value $(x_1, x_2) = (F^{-1}_{\chi^2_{10}}(\Phi(4)), F^{-1}_{\chi^2_{10}}(\Phi(4)))$, where $F^{-1}_{\chi^2_{10}}(\cdot)$ is the quantile function of a χ^2_{10} distribution and Φ is the standard normal cumulative distribution function. The outliers generated in this way for the χ^2_{10} distribution are equivalent to outliers located at 4 on the scale of the normal distribution. In Figure 1, we present the scatter plot and the estimated density of the global outlyingness measures ASO_i for the clean setup as well as for the contaminated setup. For the estimated densities, we superimpose the density estimated using the transformation related to the Tukey *g*-and-*h* distribution — the so-called Tukey-based density — and a standard kernel density. A vertical line is drawn at the cut-off point (i.e., $B_+(0.01)$ in this example). The points identified as outliers by the proposed methodology are represented by hollow symbols. The cluster of generated outliers is represented by a large square in the scatter plot and is easily identifiable in the kernel density as a bump in the right side of the distribution.

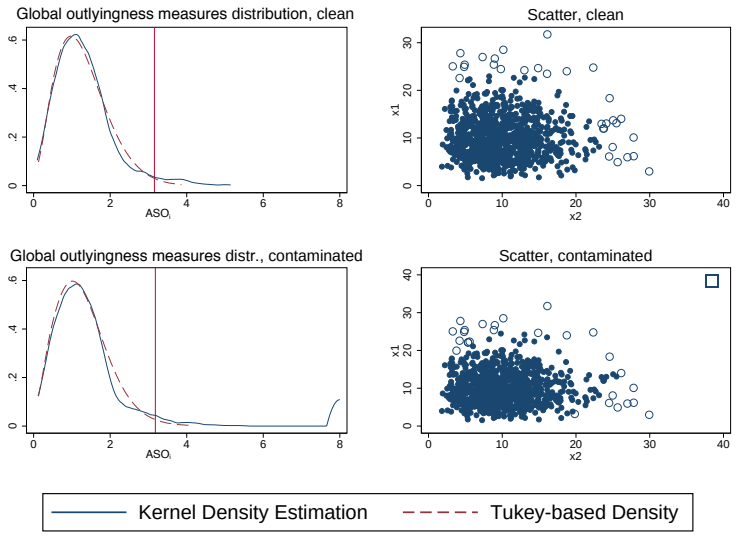

FIGURE 1. *(Right side) Scatter plots of the clean and of the contaminated bivariate samples (the points identified as outliers are represented by hollow symbols; the cluster of generated outliers is represented by a large square) – (Left side) Kernel and Tukey-based estimations of the density of the asymmetrical outlyingness measures* ASO_i *(*$i = 1, \ldots, 1000$*) in the clean and in the contaminated bivariate sample*

For the clean data, the density of the global outlyingness measures is well approximated by our methodology as can be seen when comparing the kernel density and the density estimated relying on the Tukey *g*-and-*h* transformation. As expected when using the 99th percentile cut-off, approximately 1% of the individuals are identified as outliers.

 ISSN: 2102-6238

When we concentrate on the contaminated case, the methodology correctly identifies all the generated outliers. Some standard individuals are identified as outliers as well. This was also to be expected. Indeed, since we use robust estimators for the parameters of the Tukey distribution — estimators that have a breakdown point equal to 10% — the Tukey-based estimated density of the asymmetrical outlyingness measures in the contaminated case is quite similar as in the non contaminated case. It is only slightly more skewed and heavy-tailed than when no outlier is present, which implies than the cut-off line moves very slightly to the right. Hence, the percentage of standard observations spotted as atypical is only slightly lower in the contaminated case compared to the clean case. Due to the robustness properties of the estimators for the parameters of the Tukey distribution, this latter feature is not related to the degree of outlyingness of the outliers that contaminate the data set.

A more detailed analysis of the sensitivity — the percentage of outliers correctly identified as atypical observations — and the specificity — one minus the percentage of non outlying observations erroneously identified as outliers — of the method will take place in the simulations section (see Section 6).

5.2. *Quality of the adjustment of the distribution of the w_i's ($i = 1,\ldots,n$) by a Tukey distribution*

In Section 4, we argue that the Tukey g-and-h distribution allows to adjust adequately the distribution of the transformed asymmetrical outlyingness measures w_i ($i = 1,\ldots,n$) regardless of the skewness and the tails heaviness of the multivariate distribution of the data $\mathbf{x}_i$ ($i = 1,\ldots,n$). The only restriction that we have to impose to the underlying multivariate distribution of the original data is unimodality in order to ensure the unimodality of the distribution of the outlyingness measures.

To illustrate the quality of the adjustment of the distribution of the w_i's by a Tukey distribution, we have generated six bivariate samples $\{(x_{i1},x_{i2}); i = 1,\ldots,n\}$ of size $n = 1000$ using six very different distributions [9]: the standard normal distribution, the Student distribution with 2 degrees of freedom (that is symmetrical with very heavy tails), the Exponential distribution with rate parameter equal to one (that is skewed with a moderate tail heaviness), the Fréchet distribution with shape parameter equal to 2 (that is skewed with severe tail heaviness), the Triangular distribution with support $[0,1]$ and mode at 0.1 (that is skewed with bounded support and with a corner point at 0.1), and the Beta distribution with shape parameters equal to 2 and 5 (that is skewed, defined on the interval $[0,1]$ and smooth everywhere) [10].

Figure 2 presents, for each of the six cases, the histogram of the transformed asymmetrical outlyingness measures w_i ($i = 1,\ldots,1000$) on which is superposed the density of the estimated Tukey distribution used to adjust the observed distribution of the w_i's; it also indicates the p-value associated with the Kolmogorov-Smirnov goodness-of-fit test. All the p-values are greater or equal to 0.28 and lead us to not reject the null hypothesis according to which the w_i's follow a Tukey distribution.

[9] As in the previous illustrative example (see Section 5.1), the values x_{i1} ($i = 1,\ldots,1000$) have been generated independently of the values x_{i2} ($i = 1,\ldots,1000$).

[10] The graphs of the univariate and bivariate densities considered here are available in Appendix 2.

 ISSN: 2102-6238

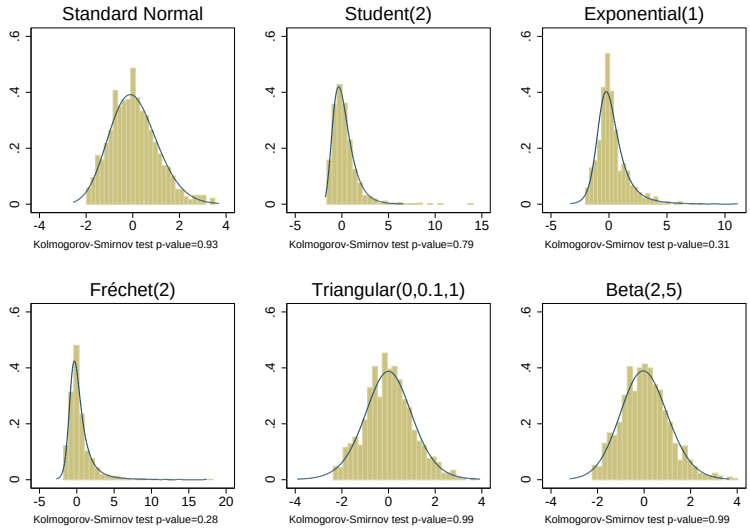

FIGURE 2. *Histogram of the transformed asymmetrical outlyingness measures w_i ($i = 1,\ldots,1000$), density of the estimated Tukey distribution adjusting the observed distribution of the w_i's, and p-value of the Kolmogorov-Smirnov goodness-of-fit test for a bivariate data set $\{(x_{i1},x_{i2}); i = 1,\ldots,1000\}$ generated from a (i) standard normal, (ii) Student(2), (iii) Exponential(1), (iv) Fréchet(2), (v) Triangular(0,0.1,1), (vi) Beta(2,5)*

Other multivariate distributions of the data have been tested. See, for instance, the simulations discussed in Section 6.2.

6. Simulations

6.1. *First set of simulations*

In the first set of simulations, we generate $B = 1000$ multivariate samples of size n and of dimension p, from each of the six distributions considered in Section 5.2. Then, we contaminate the samples by replacing a certain proportion ε of data by outlying values. We consider two sample sizes ($n = 100, 1000$), three different dimensions ($p = 2, 5, 10$) and three contamination levels ($\varepsilon = 0\%, 1\%, 5\%$).

The data are actually generated in the following way. For b from 1 to $B = 1000$:

1. We generate p independent pseudo-random samples $\{z_{ik}^{(b)}; i = 1,\ldots,n\}$ $(k = 1,\ldots,p)$ from the $N(0,1)$ distribution.
2. Denoting by Φ the cumulative distribution function (cdf) of the standard normal distribution and by F the cdf of the target distribution, $\{\Phi(z_{ik}^{(b)}); i = 1,\ldots,n\}$ $(k = 1,\ldots,p)$ are p independent pseudo-random samples from the uniform $U(0,1)$ distribution and, hence, $\{\mathbf{x}_i^{(b)} = \left(x_{i1}^{(b)},\ldots,x_{ip}^{(b)}\right)^t; i = 1,\ldots,n\}$ with $x_{ik}^{(b)} = F^{-1}\left(\Phi(z_{ik}^{(b)})\right)$ $(k = 1,\ldots,p)$ is a pseudo-random p-multivariate sample from the target distribution.

 ISSN: 2102-6238

3. We contaminate the obtained sample $\{\mathbf{x}_i^{(b)}; i = 1, \ldots, n\}$ by replacing a proportion ε of randomly selected $\mathbf{x}_i^{(b)}$ with $\mathbf{x}_{i;\text{out}}^{(b)} = \left(F^{-1}(\Phi(4)), \ldots, F^{-1}(\Phi(4))\right)^t$. By generating outliers in this way, the first four cases — standard normal, Student(2), Exponential(1) and Fréchet(2) — can be compared in terms of average sensitivity and specificity of the outlier detection method since the degree of outlyingness in the contamination is the same (around the value 4 on the scale of the standard normal). For the Triangular and Beta distributions which have a compact support on $[0, 1]$, outliers are generated slightly differently in such a way that they lie outside of the support; we simply take $\mathbf{x}_{i;\text{out}}^{(b)}$ as the vector having each of its p components equal to $\mu + 4\sigma$, where the theoretical mean μ and standard deviation σ are respectively 0.3291 and 0.2248 for the Triangular(0,0.1,1) distribution, and 0.2645 and 0.1597 for the Beta(2,5) distribution.

Table 1 reports the average sensitivity — the average proportion of outliers detected by the method as atypical points in the data set — and the average specificity — one minus the average proportion of non outlying points erroneously detected as outliers by the method — over the 1000 replications, for each specification of the sample size n, of the dimension p and of the percentage of contamination ε.

The results of the simulations clearly point towards the good behavior of the methodology. Both specificity and sensitivity are very good in small samples. In large samples the performance of the methodology improves and reaches a sensitivity of almost 100% for all simulations and contamination setups, and a specificity of approximately 99% (as expected, since we consider a cut-off value $\xi_{1-\alpha}$ with $1 - \alpha = 0.99$).

6.2. *Second set of simulations*

In order to verify if the proposed outlier identification method works well whatever are the skewness and the tail weight of the data distribution, we run a second set of simulations in which we consider the family of the SAS-normal [11] distributions defined as follows (see Jones and Pewsey, 2009): if Z is a random variable with standard normal distribution, and η and δ are two constants ($\eta \in \mathbb{R}, \delta > 0$), then the random variable Y given by

$$Y = \sinh\left[\frac{1}{\delta}\left(\sinh^{-1}(Z) + \eta\right)\right]$$

has a SAS-normal(η, δ) distribution [12].

For values of the skewness parameter η and of the shape parameter δ ranging (by step of 0.1) from -1 to 1 and from 0.2 to 2, respectively, we first generate $B = 1000$ replications of a series of $n = 1000$ observations from a 2-dimensional (independent) SAS-normal(η, δ) distribution. In each of the B series, we then replace 5% of randomly selected observations by outliers located (i)

[11] SAS: sinh-arcsinh.

[12] To have an idea of how the SAS-normal distribution behaves with respect to changes in its parameters η and δ, the reader may refer to Appendix 3: In this appendix, some densities of the SAS-normal corresponding to various values of η and δ are represented. The bottom right graph presents the most extreme distributions considered in the simulation setup of Section 6.2.

 ISSN: 2102-6238

TABLE 1. *Average sensitivity and average specificity (in percentage) over the* 1000 *replications related to the detection of outliers by the new method*

	p	ε	Sensitivity		Specificity	
			$n = 100$	$n = 1000$	$n = 100$	$n = 1000$
$N(0,1)$	2	0%	–	–	97.1	98.3
	5	0%	–	–	98.0	97.7
	10	0%	–	–	99.3	99.1
	2	1%	100	100	97.1	98.6
	5	1%	100	100	97.4	98.8
	10	1%	100	100	97.6	98.8
	2	5%	98.1	100	97.2	98.6
	5	5%	99.6	100	97.5	98.7
	10	5%	97.7	100	97.7	98.7
t_2	2	0%	–	–	97.0	98.7
	5	0%	–	–	98.0	98.3
	10	0%	–	–	98.3	99.0
	2	1%	100	100	97.5	98.5
	5	1%	99.9	100	97.5	98.6
	10	1%	100	100	97.9	98.7
	2	5%	100	100	97.5	98.6
	5	5%	99.9	100	97.5	98.7
	10	5%	99.8	100	97.9	98.7
$Exp(1)$	2	0%	–	–	99.2	98.2
	5	0%	–	–	98.9	99.1
	10	0%	–	–	99.1	99.0
	2	1%	100	100	97.8	99.9
	5	1%	100	100	98.0	99.2
	10	1%	100	100	97.8	98.5
	2	5%	99.3	96.9	98.7	99.9
	5	5%	100	100	97.9	98.8
	10	5%	99.9	100	97.4	98.1
Fréchet(2)	2	0%	–	–	96.2	99.3
	5	0%	–	–	97.0	97.9
	10	0%	–	–	98.0	98.3
	2	1%	100	100	98.9	99.8
	5	1%	100	100	97.4	97.9
	10	1%	100	100	97.5	98.3
	2	5%	100	98	98.9	99.8
	5	5%	100	100	97.4	97.8
	10	5%	99.6	100	97.5	98.6
Triangular(0,0.1,1)	2	0%	–	–	99.0	98.8
	5	0%	–	–	99.0	99.3
	10	0%	–	–	99.1	99.1
	2	1%	100	100	97.1	98.7
	5	1%	100	100	97.5	99.3
	10	1%	100	100	97.7	99.2
	2	5%	99.1	100	98.6	99.9
	5	5%	100	100	98.7	99.9
	10	5%	100	100	98.9	99.9
Beta(2,5)	2	0%	–	–	99.0	98.9
	5	0%	–	–	98.2	99.3
	10	0%	–	–	98.3	99.0
	2	1%	100	100	97.0	98.7
	5	1%	100	100	97.7	99.3
	10	1%	100	100	97.8	99.2
	2	5%	92.2	99.7	98.6	99.6
	5	5%	97.4	100	98.7	99.9
	10	5%	94.5	100	98.8	99.9

 ISSN: 2102-6238

at the value of 4 and (ii) at the value of 5 on the scale of the standard normal [13]. For each value of η and δ and for each of both values of contamination, the average sensitivity over the $B = 1000$ replications is determined. The results are synthesized in Figure 3 using contour maps.

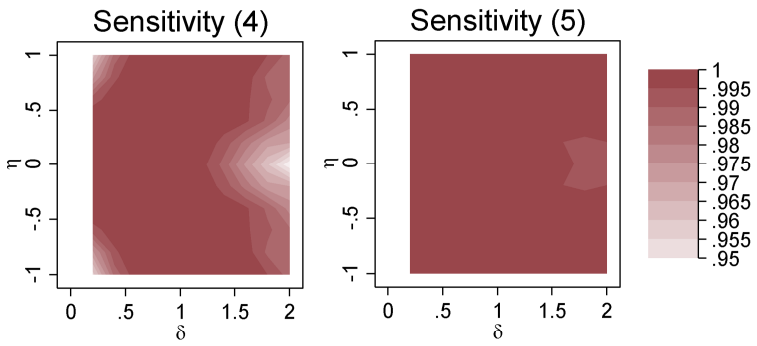

FIGURE 3. *Average sensitivity of the outlier identification method in case of bivariate samples of size* $n = 1000$ *from 2-dimensional (independent) SAS-normal*(η, δ) *where* $\eta \in [-1, 1]$ *and* $\delta \in [0.2, 2]$*, with 5% of outliers at the value of 4 (left figure) or at the value of 5 (right figure) on the scale of the standard normal*

As can be seen, the sensitivity of the method is very high even for very extreme distributions of the data.

6.3. *Third set of simulations*

The third set of simulations has for objective to compare our outlier detection method with its closest competitor — the method of Hubert and Van der Veeken (2008) — in terms of sensitivity as well as in terms of computational complexity and time.

We actually repeat a simulation setup similar as the one presented in Section 6.1 with $n = 1000$ and a percentage ε of contamination equal to 5%, but we consider here outliers at values ranging (by step of 0.5) from 0 to 5.5 on the scale of the standard normal.

As mentioned in Section 2.2, the outlier detection method of Hubert and Van der Veeken is once for all calibrated in such a way that, in absence of contamination, about 0.35% ($= 0.7\%/2$) of the observations will be identified as atypical. In the method proposed here, we have the freedom to fix ourselves the expected percentage of observations in a clean data set that will be considered as outliers: This expected percentage is equal to α if we decide to use the $(1-\alpha)$-quantile $\xi_{1-\alpha}$ of the $T_{\widehat{g},\widehat{h}}(\widehat{A},\widehat{B})$ distribution in Step 4 of the procedure (see Section 4.2). This freedom is very interesting, especially when we suspect that the data set is contaminated by mild outliers [14].

[13] For $b = 1, \ldots, B$, we replace 5% of randomly selected observations $\mathbf{x}_i^{(b)}$ by $\mathbf{x}_{i;\text{out}}^{(b)} = \left(F^{-1}(\Phi(4)), F^{-1}(\Phi(4))\right)^t$ in the first setup and by $\mathbf{x}_{i;\text{out}}^{(b)} = \left(F^{-1}(\Phi(5)), F^{-1}(\Phi(5))\right)^t$ in the second setup, where F is the cdf of the SAS-normal(η, δ).

[14] Mild outliers, inspite of their low degree of outlyingness, are often dangerous since they can *inter alia* strongly distort the estimation of the covariance matrix of the multivariate distribution of the data. It is then crucial to be able to identify correctly these mild outliers before to work with the data.

 ISSN: 2102-6238

The results of the third set of simulations clearly illustrate this fact. As shown by the sensitivity curves in Figure 4, large outliers (i.e., at a value of 5 on the scale of the standard normal) are properly identified by both methods. There is however a big difference between the new method and the method of Hubert and Van der Veeken in terms of sensitivity with respect to mild outliers. Whatever is the value of p and the underlying distribution of the data, the proportion of mild outliers — consider, for example, outliers ranging from 2 to 3 on the scale of the standard normal — correctly identified as atypical is quite low with the method of Hubert and Van der Veeken, and appears much higher with the new method when we take $\alpha = 1\%$. If we consider $\alpha = 5\%$, the sensitivity curves are translated to the left with respect to the case of $\alpha = 1\%$, leading to a very high sensitivity even when the outliers are located between 2 and 2.5 on the scale of the standard normal.

The new outlier detection method proposed in this paper presents another advantage with respect to the method of Hubert and Van der Veeken: It has a lower computational complexity, equal to $O(np)$, while the method of Hubert and Van der Veeken possesses a computational complexity of $O(np\log n)$. This difference in computational complexity induces a significant difference in terms of computing time.

To illustrate the difference of time performances of both methods, we have decided to compare the average time — over 1000 replications — required to run both of outlier identification methods. The method of Hubert and Van der Veeken, based on the computation of the *adjusted* global outlyingness measures AO_i $(i = 1,\ldots,n)$, has been implemented with the open and very efficient R code "adjOutlyingness". For the new method, based on the computation of the *asymmetrical* global outlyingness measures ASO_i $(i = 1,\ldots,n)$, we have used our own R code. In both cases, we have determined the global outlyingness measures on the basis of the projection of the observations on $250p$ directions of the space. All simulations have been performed using R 3.2.3 called from Stata 13.1 on a PC Intel(R) Core(TM) i7-4770 CPU@3.40Ghz, 8Gb of RAM.

Table 2 presents the average computing times in seconds — for the AO-method of Hubert and Van der Veeken in the upper half of each cell and for the new ASO-method in the lower half of each cell — obtained for various sample sizes n and various dimensions p. When np increases, the difference in speed between the two methods becomes very clear.

TABLE 2. *Average (over 1000 replications) computing times in seconds for the AO-method of Hubert and Van der Veeken (in the upper half of each cell) and for the ASO-method (in the lower half of each cell) for various sample sizes n and various dimensions p*

p \ n	1000	5000	10000	50000
2	0.44 **0.81**	2.17 **1.45**	4.40 **2.12**	36.02 **7.33**
5	1.06 **1.37**	5.30 **2.65**	11.13 **4.04**	63.90 **17.13**
10	2.14 **3.20**	10.64 **4.84**	23.81 **7.96**	148.1 **41.39**
25	5.16 **5.04**	35.65 **13.17**	57.23 **21.15**	309.6 **96.20**

 ISSN: 2102-6238

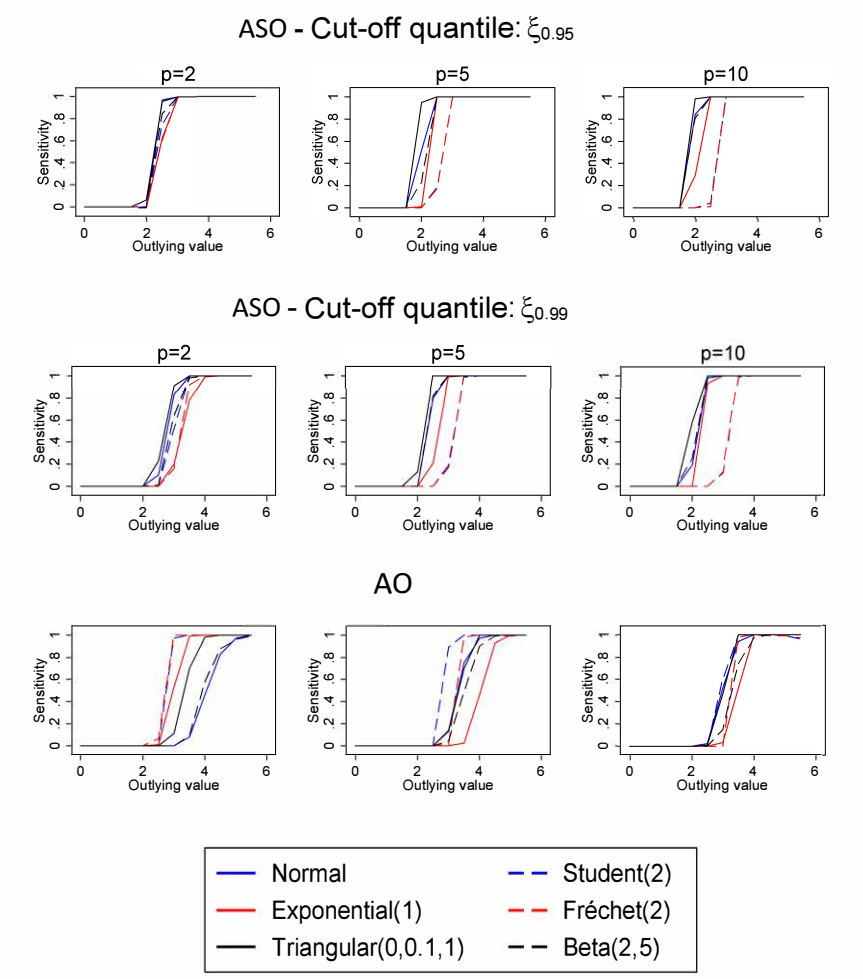

FIGURE 4. *Average sensitivity curves over the 1000 replications, for $\varepsilon = 5\%$ of contamination of the sample of size $n = 1000$ and of dimension $p = 2$, 5 or 10, with the outliers at values ranging from 0 to 5.5 on the scale of the standard normal. (At the top) ASO, $\alpha = 5\%$: New outlier identification method based on the asymmetrical outlyingness measures ASO_i, with $\xi_{1-\alpha} = \xi_{0.95}$ – (In the middle) ASO, $\alpha = 1\%$: Idem, with $\xi_{1-\alpha} = \xi_{0.99}$ – (At the bottom) AO: Outlier identification method of Hubert and Van der Veeken, based on the adjusted outlyingness measures AO_i*

7. Application

In this application, we try to identify outlying counties in the US in terms of the relation obesity-physical inactivity-diabetes for the year 2010. The data are available from the Centers for Disease Control and Prevention (CDC) at http://www.cdc.gov/diabetes/atlas/countydata/County_ListofIndicators.html. These data correspond to the percentage of diabetes, obesity and physical inactivity prevalence in all US counties (except for 3 missing observations that we do not take into account for our analysis). The size of the analysed sample is therefore $n = 3143$ and the number p of dimensions is equal to 3.

We use the methodology discussed above. We first determine the asymmetrical global outlying-

 ISSN: 2102-6238

ness measures ASO_i ($i = 1, \ldots, 3143$). In Figure 5, we present the kernel estimation as well as the Tukey-based estimation of the density of these measures. We draw a line at percentile 99 of the latter distribution (i.e. 3.58) to illustrate the cut-off point.

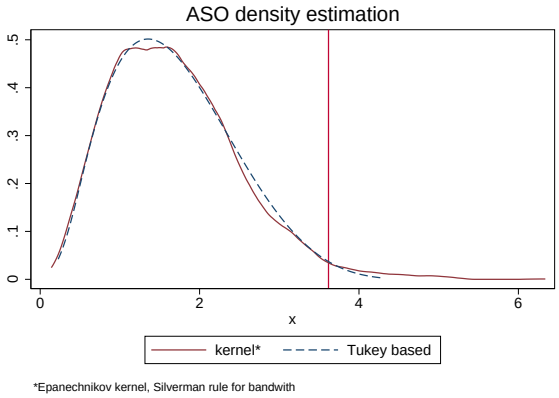

FIGURE 5. *Kernel and Tukey-based estimations of the density of the asymmetrical outlyingness measures* ASO_i *for the data of the CDC* ($n = 3143$, $p = 3$)

Using this cut-off point (with $\alpha = 1\%$), 72 counties are identified as outlying (approximately 2.3% of the counties). Interestingly, several types of outliers seem to be present. Firstly, some counties are extreme in the sense that they have very high levels in the three dimensions. This is for example the case for Greene county in Alabama or Claiborne county in Mississippi. These counties have respectively 36.5% and 39.7% prevalence of inactivity, 19.7% and 16.1% prevalence of diabetes and 47.9% and 41.4% prevalence of obesity. On the other extreme other counties such as Denver county and Ouray county in Colorado have very low levels in the three dimensions. These counties have respectively 14.3% and 14.8% prevalence of inactivity, 6.1% and 6.4% prevalence of diabetes and 18.2% and 17.5% prevalence of obesity. Then you have other cases such as Potter county in South Dakota or Johnson county in Nebraska that have relatively high levels of physical inactivity (respectively 29.6% and 32.9%) and obesity (respectively 28.4% and 30%) but relatively low levels of diabetes prevalence (8.7% and 8.2%).

If we compare the results with those obtained by a standard Stahel-Donoho estimator that assumes elliptical symmetry, we have that 70 of the 72 counties identified as outliers by the methodology proposed here are identified as such by the standard method. On the other hand, the standard methodology identifies 25 individuals as outliers that our methodology identifies as standard.

8. Conclusion

In multivariate analysis, it is fairly difficult to identify outliers in case of skewed or heavy-tailed distributions. Indeed, most commonly used identification tools such as robust Mahalanobis distances rely on the elliptical symmetry assumption. Recently, some authors such as Hubert and Van der Veeken (2008) have proposed some tools for multivariate outlier identification but these are associated to very large computational complexities and cope mainly with skewed and not

 ISSN: 2102-6238

heavy-tailed distributions and unbounded supports. In this paper we propose a very simple method based on projections that keeps the computational complexity of the problem very low, of the order $O(np)$. The proposed method works well with both asymmetrical and/or heavy-tailed unimodal distributions, and with both bounded and unbounded supports. We run several simulations and the new method seems to perform well for a very wide range of single peaked distributions.

References

Bruffaerts, C., Verardi, V., and Vermandele, C. (2014). A generalized boxplot for skewed and heavy-tailed distributions. *Statistics and Probability Letters*, 95(1):110–117.

Brys, G., Hubert, M., and Struyf, A. (2004). A robust measure of skewness. *Journal of Computational and Graphical Statistics*, 13(4):996–1017.

Donoho, D. (1982). Breakdown properties of multivariate location estimators. Technical report, Harvard University, Boston. Qualifying paper.

Hoaglin, D., Mosteller, F., and Tukey, J. (1985). *Exploring Data Tables, Trends and Shapes*. Wiley, New York.

Hubert, M. and Van der Veeken, S. (2008). Outlier detection for skewed data. *J. Chemometrics*, 22(3-4):235–246.

Hubert, M. and Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Comput. Stat. Data Anal.*, 52(12):5186–5201.

Jiménez, J. and Arunachalam, V. (2011). Using Tukey's g and h family of distributions to calculate value-at-risk and conditional value-at-risk. *J. Risk*, 13(4):95–116.

Jones, M. C. and Pewsey, A. (2009). Sinh-arcsinh distributions. *Biometrika*, 96(4):761–780.

Ley, C. (2015). Flexible modelling in statistics: past, present and future. *Journal de la Société Française de Statistique*, 156(1):76–96.

MacGillivray, H. (1992). Shape properties of the g-and-h and Johnson families. *Communications in Statistics - Theory and Methods*, 21(5):1233–1250.

Mahbubul, A., Majumder, A., and Ali, M. (2008). A comparison of methods of estimation of parameters of Tukey's gh family of distributions. *Pakistan Journal of Statistics*, 24(2):135–144.

Maronna, R. and Yohai, V. (1995). The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90(429):330–341.

Martinez, J. and Iglewicz, B. (1984). Some properties of the Tukey g-and-h family of distributions. *Communications in Statistics - Theory and Methods*, 13(3):353–369.

Rousseeuw, P. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283.

Stahel, W. (1981). *Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen*. PhD thesis, ETH Zürich.

Tukey, J. (1977). Modern techniques in data analysis. In *Proceedings of the NSF-Sponsored Regional Research Conference*.

Xu, G. and Genton, M. (2015). Efficient maximum approximated likelihood inference for Tukey's. *Computational Statistics & Data Analysis*, 91:78–91.

Xu, Y., Iglewicz, B., and Chervoneva, I. (2014). Robust estimation of the parameters of g-and-h distributions, with application to outlier detection. *Computational Statistics & Data Analysis*, 75:66–80.

Appendix 1: The estimation of the parameters of the $T_{g,h}(A,B)$ distribution

Let $Z \sim N(0,1)$, $Y = A + B\tau_{g,h}(Z) \sim T_{g,h}(A,B)$ and $\mathscr{Y}^{(n)} = \{y_1, \ldots, y_n\}$ be a series of n independent realizations of Y. For $\nu \in (0,1)$, let us denote by z_ν and y_ν the quantiles of order ν of the $N(0,1)$ and of the $T_{g,h}(A,B)$ distributions, respectively, and by $Q_\nu(\mathscr{Y}^{(n)})$ the empirical quantile of order ν in the series $\mathscr{Y}^{(n)}$.

 ISSN: 2102-6238

a) Estimation of A

Since $\tau_{g,h}(z)$ is a one-to-one monotone function of $z \in \mathbb{R}$, we have, for all $\nu \in (0,1)$:

$$y_\nu = A + B\tau_{g,h}(z_\nu). \tag{8}$$

Since $z_{0.5} = 0$ and $\tau_{g,h}(0) = 0$, relation (8) implies that $y_{0.5} = A$. The location parameter A corresponds to the median of the Tukey distribution and can be simply estimated by the empirical median of the series $\mathscr{Y}^{(n)}$:

$$\widehat{A} = Q_{0.5}(\mathscr{Y}^{(n)}). \tag{9}$$

b) Estimation of g

In the outlier identification method we propose, we exclusively have to adjust right-skewed distributions by Tukey g-and-h distributions. Hence, we may here restrict ourselves to the case where g has a non-zero value and is strictly positive.

Relation (8) actually implies, for $\nu > 0.5$, that

$$\frac{\mathrm{UHS}_\nu}{\mathrm{LHS}_\nu} = \exp(gz_\nu)$$

where UHS_ν and LHS_ν are the ν-th upper and lower half spreads of the $T_{g,h}(A,B)$ distribution, respectively:

$$\mathrm{UHS}_\nu = y_\nu - y_{0.5},$$
$$\mathrm{LHS}_\nu = y_{0.5} - y_{1-\nu}.$$

Consequently, for any $\nu > 0.5$:

$$g = \frac{1}{z_\nu} \ln\left(\frac{\mathrm{UHS}_\nu}{\mathrm{LHS}_\nu}\right).$$

A natural estimate of the parameter g is then given by

$$\widehat{g}_\nu = \frac{1}{z_\nu} \ln\left(\frac{\mathrm{UHS}_\nu(\mathscr{Y}^{(n)})}{\mathrm{LHS}_\nu(\mathscr{Y}^{(n)})}\right) \tag{10}$$

for any fixed order $\nu \in (0.5, 1)$, where

$$\mathrm{UHS}_\nu(\mathscr{Y}^{(n)}) = Q_\nu(\mathscr{Y}^{(n)}) - Q_{0.5}(\mathscr{Y}^{(n)}),$$
$$\mathrm{LHS}_\nu(\mathscr{Y}^{(n)}) = Q_{0.5}(\mathscr{Y}^{(n)}) - Q_{1-\nu}(\mathscr{Y}^{(n)}).$$

This estimator of g has a breakdown point of $(1-\nu)\times 100\%$. From then on, by choosing $\nu = 0.90$, we make so that the breakdown point of the estimator of g is as high as 10%.

Naturally, if one wants to increase the efficiency of the estimator of g by using more information contained in the data, and especially in the tails of the observed distribution of $\mathscr{Y}^{(n)}$, several values of ν could be considered and g could be estimated by the median of the estimates $\widehat{g}_\nu$ associated with these different values of $\nu \in (0.5, 1)$ (see Jiménez and Arunachalam, 2011). But simulations showed us that, in the context of estimation of the parameters of a Tukey distribution considered in this paper, it is not necessary to use this alternative estimator of g since we already get very good results when simply using $\widehat{g}_{0.90}$.

 ISSN: 2102-6238

c) Estimation of B and h

When $g \neq 0$, the parameter that controls the elongation of the tails (h) and the scale parameter (B) can be jointly estimated conditionally on the value of g. Indeed, as highlighted by Hoaglin et al. (1985),

$$\ln(y_{0.5} - \theta_v) = \ln\left(\frac{B}{g}\right) + \frac{1}{2}hz_v^2 \tag{11}$$

with $\theta_v < y_{0.5}$ for all $v > 0.5$ and, for all $v \in (0,1)$, $v \neq 0.5$,

$$\theta_v = \frac{y_v y_{1-v} - y_{0.5}^2}{\mathrm{UHS}_v - \mathrm{LHS}_v}\ .$$

Note that $\theta_v = \theta_{1-v}$. The values of B and h can then be estimated using a linear regression of $\ln\left(Q_{0.5}(\mathscr{Y}^{(n)}) - \theta_v^{(n)}\right)$ on $\frac{1}{2}z_v^2$, for different values of v, where $\theta_v^{(n)}$ is the empirical counterpart of θ_v: The estimate of h corresponds in that case to the estimate of the slope parameter and the value of the scale parameter B is estimated by multiplying the previously estimated value of g by the exponential of the estimate of the intercept term (see Jiménez and Arunachalam, 2011).

We propose here a quite different procedure in which h and B are estimated separately.

c.1) Estimation of B

If $Y = A + B\tau_{g,h}(Z) \sim T_{g,h}(A,B)$, then $Y^* = \frac{Y-A}{B} = \tau_{g,h}(Z) \sim T_{g,h}(0,1)$. Denoting by y_v^* ($v \in (0,1)$) the quantile of order v of the standardized $T_{g,h}(0,1)$ distribution and defining $\mathrm{IQR} = y_{0.75} - y_{0.25}$ and $\mathrm{IQR}^* = y_{0.75}^* - y_{0.25}^*$ — IQR and IQR* are the interquartile ranges of the $T_{g,h}(A,B)$ and the $T_{g,h}(0,1)$ distributions, respectively — we clearly have:

$$\mathrm{IQR} = B\ \mathrm{IQR}^*$$

and hence:

$$B = \frac{\mathrm{IQR}}{\mathrm{IQR}^*} = \frac{c\mathrm{IQR}}{c\mathrm{IQR}^*} \tag{12}$$

for $c = 1/(z_{0.75} - z_{0.25}) = 0.7413$ (c is the constant factor ensuring that the empirical interquartile range is a Fisher consistent estimator of the scale parameter σ in the Gaussian case).

Extensive simulations have allowed us to establish that there exists an almost perfect relation between $c\mathrm{IQR}^*$ on the one hand, and, on the other hand, the quantiles-based measures of skewness

$$\mathrm{SK} = \frac{y_{0.90}^* + y_{0.10}^* - 2y_{0.5}^*}{y_{0.90}^* - y_{0.10}^*} = \frac{y_{0.90} + y_{0.10} - 2y_{0.5}}{y_{0.90} - y_{0.10}}$$

and of kurtosis

$$\mathrm{T} = \frac{y_{0.90}^* - y_{0.10}^*}{y_{0.75}^* - y_{0.25}^*} = \frac{y_{0.90} - y_{0.10}}{y_{0.75} - y_{0.25}}\ .$$

 ISSN: 2102-6238

We have [15]:

$$c\text{IQR}^* \simeq 0.6817766 + 0.0534282\ \text{SK} + 0.1794771\ \text{T} - 0.0059595\ \text{T}^2. \tag{13}$$

Hence, a natural estimate of B is:

$$\widehat{B} = \frac{c\text{IQR}(\mathscr{Y}^{(n)})}{\widehat{c\text{IQR}^*}}, \tag{14}$$

where $\widehat{c\ \text{IQR}^*}$ is obtained by replacing, in the right member of (13), SK and T by their empirical counterparts $\text{SK}(\mathscr{Y}^{(n)})$ and $\text{T}(\mathscr{Y}^{(n)})$, respectively.

c.2) Estimation of h

Let us consider consider once again $Y^* = \frac{Y-A}{B}$, where $Y \sim T_{g,h}(A,B)$:

$$Y^* \sim T_{g,h}(0,1).$$

Then, relation (11) applied for Y^* becomes:

$$\ln(-\theta_\nu^*) = \ln\left(\frac{1}{g}\right) + \frac{1}{2}hz_\nu^2$$

for all $\nu > 0.5$, where

$$\theta_\nu^* = \frac{y_\nu^* y_{1-\nu}^* - (y_{0.5}^*)^2}{\text{UHS}_\nu^* - \text{LHS}_\nu^*} = \frac{y_\nu^* y_{1-\nu}^*}{y_\nu^* + y_{1-\nu}^*}\ .$$

Consequently, for all $\nu > 0.5$:

$$h = \frac{2}{z_\nu^2}\ln(-g\theta_\nu^*).$$

We then obtain a natural estimate of h by taking, for any fixed $\nu > 0.5$,

$$\widehat{h}_\nu = \frac{2}{z_\nu^2}\ln(-\widehat{g}_\nu \theta_\nu^{*(n)}) \tag{15}$$

where $\widehat{g}_\nu$ is the estimate of g and $\theta_\nu^{*(n)}$ is the empirical counterpart of θ_ν^* associated with the series $\mathscr{Y}^{*(n)} = \{y_1^*, \ldots, y_n^*\}$ where $y_i^* = (y_i - \widehat{A})/\widehat{B}$ $(i = 1, \ldots, n)$.

Once again, we choose $\nu = 0.90$ in order to have a breakdown point of 10% for the estimator of h.

[15] For each value of g varying from 0.001 to 2 (by step of 0.01) and each value of h varying from -0.2 to 2 (by step of 0.01), we have generated a sample $\mathscr{Y}_{g,h}^{*(n)}$ of size $n = 1000$ from a standardized $T_{g,h}(0,1)$ distribution. In each of these samples — we actually had 43780 samples — we have determined the (corrected) interquartile range $c\text{IQR}(\mathscr{Y}_{g,h}^{*(n)})$, the skewness measure $\text{SK}(\mathscr{Y}_{g,h}^{*(n)})$ and the kurtosis measure $\text{T}(\mathscr{Y}_{g,h}^{*(n)})$. Using these measures, we have estimated (by ordinary least squares) the following regression model: $c\text{IQR}^* = \beta_0 + \beta_1\text{SK} + \beta_2\text{T} + \beta_3\text{T}^2 + \varepsilon$. We have obtained $\widehat{\beta}_0 = 0.6817766$, $\widehat{\beta}_1 = 0.0534282$, $\widehat{\beta}_2 = 0.1794771$ and $\widehat{\beta}_3 = -0.0059595$, with an adjusted R-squared of 0.9966. Note that we can still obtained a better fit (with an adjusted R-squared of 0.9999) by considering, as explanatory variables in the regression model, $\text{SK}, \text{SK}^2, \ldots, \text{SK}^5$ and $\text{T}, \text{T}^2, \ldots, \text{T}^5$. In practice, however, it appears sufficient to consider the more parcimonious model to obtain a very good adjustment of $c\text{IQR}^*$.

 ISSN: 2102-6238

Appendix 2: The data distributions considered in Sections 5.2 and 6.1

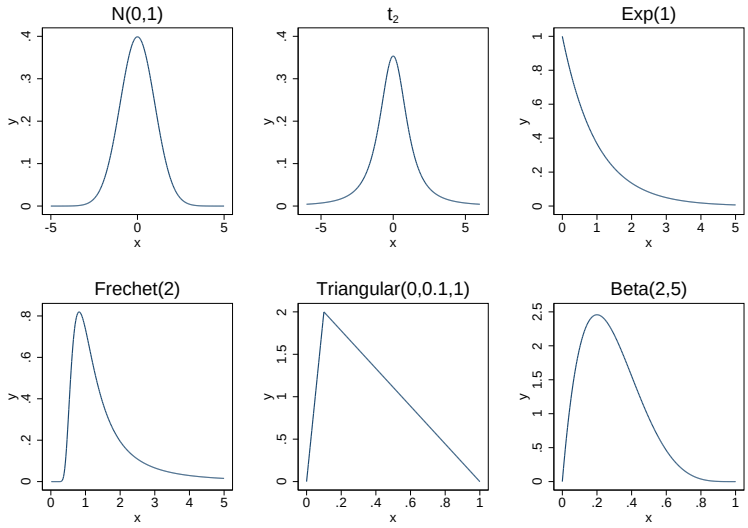

FIGURE 6. *The univariate densities considered in Sections 5.2 and 6.1*

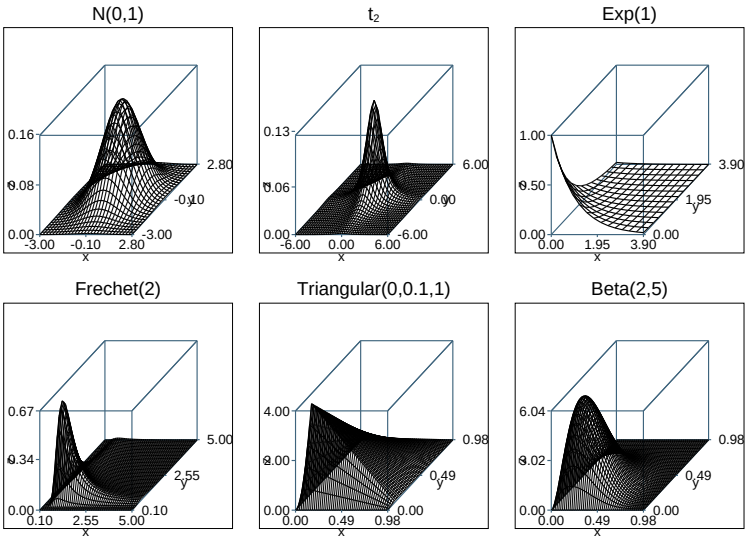

FIGURE 7. *The bivariate densities considered in Sections 5.2 and 6.1*

 ISSN: 2102-6238

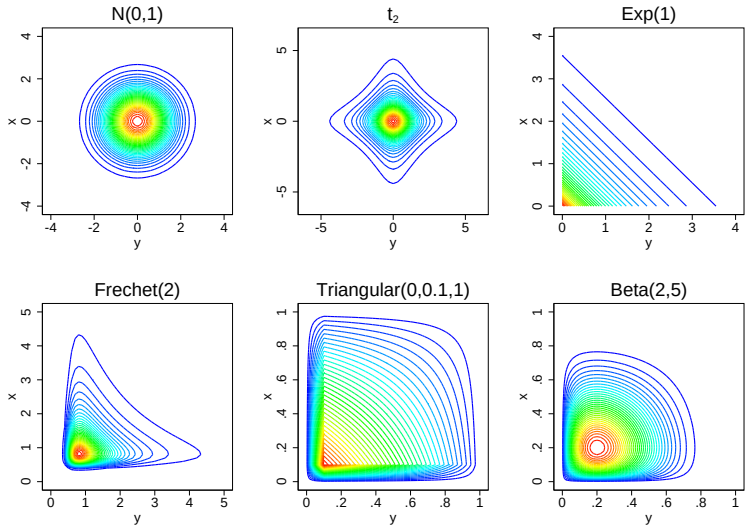

FIGURE 8. *Level curves (isocontours) for the univariate densities considered in Sections 5.2 and 6.1*

Appendix 3: The SAS-normal(η, δ) distribution

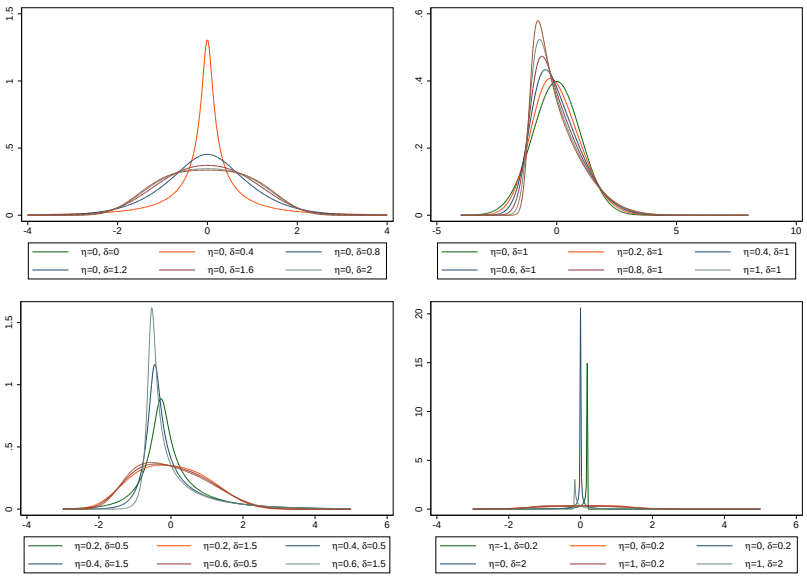

FIGURE 9. *Density function of the SAS-normal(η, δ) distribution for various values of the skewness parameter η and of the shape parameter δ*

 ISSN: 2102-6238